proteomatics

new information resources for proteomics

Proteomics is the science of understanding protein architectures at a supramolecular level, whereas the realm of understanding the organisation of biological information in its totality belongs to bioinformatics. Somewhat more prosaically, industrial bioinformatics is about data handling and analysis, while industrial proteomics uses knowledge generated by studying the proteins expressed by genomes to generate products. These may be proteins in themselves and the databases which house them at a virtual level, or may involve the use of specific, purified proteins in the development and production of a physical product. Examples of some specific products include therapeutic proteins, drug hits, drug leads and, ultimately, drugs themselves. The purpose of this article is to provide some background for those needing information on the types of discovery programmes proteomics is best placed to assist and to give some pointers to the additional informatics capabilities required to support those programmes of research. In this context we introduce the term 'proteomatics' as a means of identifying a commercial application of proteome informatics.

By Dr David J Parry-Smith and Dr David S Bailey

The central dogma of molecular biology (**Figure 1**) tells us that it is DNA that provides a template for the production of RNA, which is the message that is translated into protein – the functioning workhorse of the cell. In short, 'DNA makes RNA makes protein'. It is proteins that give cells the characteristics of their fully differentiated states. Proteins make communication systems between cells operate effectively. Ultimately, it is errors in protein function that cause disease and symptoms are expressed through the means of proteins. This analysis places proteins well and truly at the centre of the picture in terms of potential drug targets.

In recent years, the focus on genomics, driven almost entirely by technical developments, has tended to make us gather data at the DNA and RNA levels of the dogma. This information is, however, of limited value when the characteristics of the final protein product are required for target selection, expression and screening. The reason for this is the poorly understood relationship between expression at the RNA (transcriptome) level and the final form the protein product takes in its place in functioning cellular systems (proteome). Control of gene expression at the DNA level (genome), while well understood in a few cases, is not well worked out on a genomic scale.

So, we have reached a stage whereby a great deal of genomic and transcriptomic data have been collected, to a large extent simply because the technology exists to make the experiments cost-effective. They can be done on a large scale using commercially produced machinery and primed with inexpensive reagents. Once the data have been captured the informatics tools for evaluating their significance are available together with databases for storage and efficient retrieval of results.

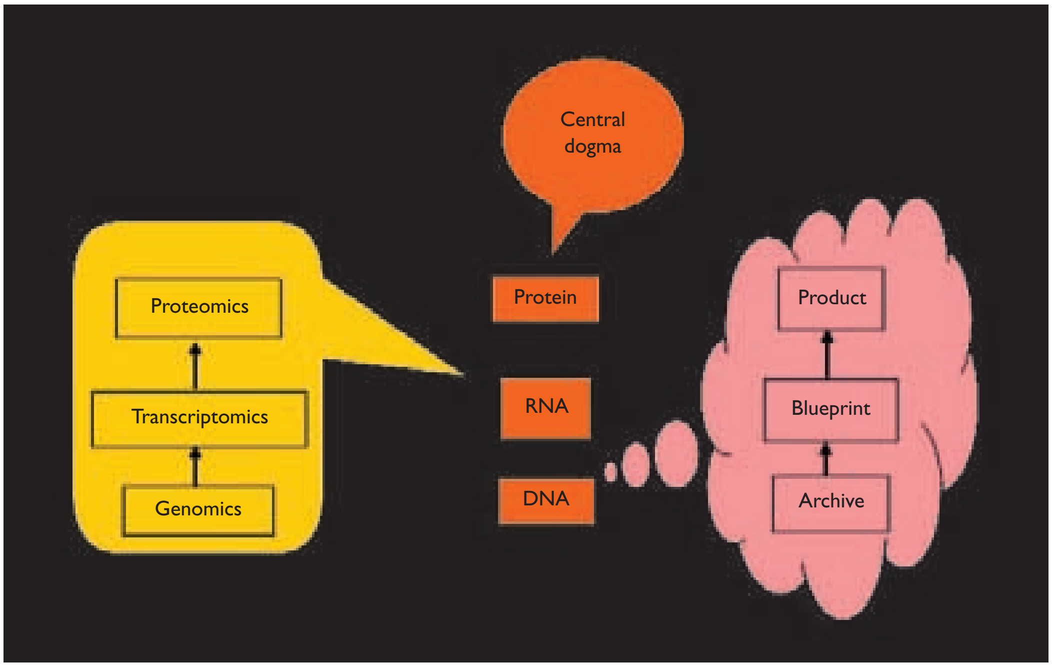


Figure 1
The central dogma of molecular biology, expressed in the red central section of the diagram, is the assertion from which the technologies of molecular biology derive their authority. The technologies of genomics, transcriptomics and proteomics, highlighted in yellow, map directly on to the three tenets of the dogma. The pink thought bubble may be a helpful tool for contemplating the three stages in the process

For proteomics, the technologies for characterisation of expressed proteins are also readily available. However, the routine production of raw material – purified proteins – at a sufficiently high level of quality and in a generic enough form to make purification of any drug discovery target in its authentic form a matter of routine (as is the case for the vast majority of genes and mRNAs encoding these targets) is not yet generally available. This is the true bottleneck in preventing us from fully exploiting proteomics as a drug discovery tool. The same bottleneck affects the supply of protein at scale for crystallisation studies that lead to structure determination by X-ray crystallography. Indeed, structural genomics is still waiting for the technology breakthrough, analogous to the introduction of automated sequencing using fluorescent dyes, that will enable the routine structural determination of proteins, both soluble and insoluble, membrane bound or not.

A brief history of proteomics

The history of proteomics stretches back to 1945 when the first sequence of the hormone insulin was worked out by Fred Sanger. **Table 1** gives a brief outline of some of the historical events on the proteomics timeline. The latest additions include cellular and chemical proteomics and proteomatics. The production of increasing volumes of data has always been a key aspect of the development of proteomics. Simply from the quantities of sequence data alone, proteomics is a large-scale science. However, when we take into consideration the comparative nature of proteomics experiments, we understand how the data volumes are expected to increase dramatically. A similar effect was seen when comparative transcript image analysis increased the size of DNA and RNA sequence repositories in the 1990s.

Definitions of proteomics

In this article we have defined proteomics (the PROTEin complement to a genOME) as the science of understanding protein architecture at a supramolecular level. This is the big picture definition. Others have defined proteomics as the qualitative and quantitative comparison of proteomes under different conditions to unravel biological processes further (ExPasy website www.expasy.ch/proteomics_def.html). Pennington and Dunn[1] take the view that proteomics is the study of protein expression and function on a genomic scale.

Whatever the words used to define the science, the technologies are varied. They meet genomics at the RNA transcript profiling interface. Genome-wide two-hybrid screening provides data on protein-protein interactions. 2-D polyacrylamide gel electrophoresis is a direct method for analysis of protein expression, sequence and structure. Mass spectrometry provides detection of proteins, gel image analysis and characterisation of proteins. Finally, informatics provides a framework for supporting and interpreting the proteomics data world.

Table 1: An outline of the development of protein sequencing leading to proteomics as a scientific discipline ready for commercial exploitation

DATE	ACHIEVEMENT	TYPE OF DATA
1945	Fred Sanger conducted pioneering work on the characterisation of Insulin by sequencing with fluorodinitrobenzene	Very low throughput protein sequence
1950	Pehr Edman developed a more efficient technique of sequential degradation from the N-terminus of the sequence	Higher throughput protein sequence
1956	Smithies and Poulik described a 2-D combination of paper and starch gel electrophoresis for the separation of serum proteins	First two dimensional co-ordinate data
1970	U.K. Laemmli ran the first 2-D PAGE for increased separation of structural proteins for a bacteriophage incorporating iso electric focusing	Production of gel image data
1975	Patrick O'Farrell published a protocol for high resolution 2-D electrophoresis	Higher quality image data impacting data storage requirements
1978	A.D. Malcolm predicted the demise of chemical sequencing of proteins in favour of nucleotide sequencing and the prediction of protein sequences	The first protein sequence data derived from predicted open reading frames
1981	Hewick and his group developed an automated protein sequencing machine Efforts are undertaken to produce an index of all human proteins	Increases data volume significantly The human protein index is begun
1986	SWISS-PROT is established as a high quality, robust protein sequence database with well researched annotation and minimal redundancy	Integrity of sequence data is improved
1989	First posters presented at symposia on the feasibility of methods for mass spectroscopic analysis of trace amounts of protein	A few labs producing quality data
1988	MALDI and ESI approaches to mass spectroscopic characterisation become available	Enhanced speed of delivery of quality protein sequence data
1993	Studies from various different groups exemplifying the new protein sequencing approaches	Low volume, high quality protein sequence data
1993	Several groups provide peptide mass searching approaches to correlate protein and gene sequence information	Improved quality of the identification of peptides
1994	The word proteomics is coined by Marc Wilkins and colleagues	
1995	The first complete sequence of a living organism (Haemophilus influenzae) is published Proteomics makes its first appearance in the published literature in July 1995	Ever higher volumes of predicted protein sequence enter the databases
1995	Correlative fragment sequence database searching linked to mass spectroscopic data yields identity of trace amount of protein in about five minutes.	High throughput protein sequencing data becomes a reality
1999	The concept of protein microarrays for analysing protein-protein interactions appears	Pathway analysis and protein expression profiles
1999	The Human Proteomics initiative is launched. This represents a major project to annotate all known human sequences according to the quality standards of SWISS-PROT, providing for each known protein the description of its function, domain structure, subcellular location, post-translational modifications, variants, similarities to other proteins, etc	A comprehensively annotated, robust, set of human sequences is envisaged
1999	Structural genomics provides a context for structure-based molecular design	Large scale architectural analysis of proteins on a genome-wide scale is initiated
2001	The Human Proteome Organization is established (www.hupo.org)	Data volumes for characterised proteins are set to expand massively
2002	Chemical proteomics first appears as a technology for focussing chemical structure space using protein architecture	Chemoinformatics data
2002	Proteomatics emerges as an integrating feature for knowledge management in proteomics	Cross-referencing of disparate proteomic data types

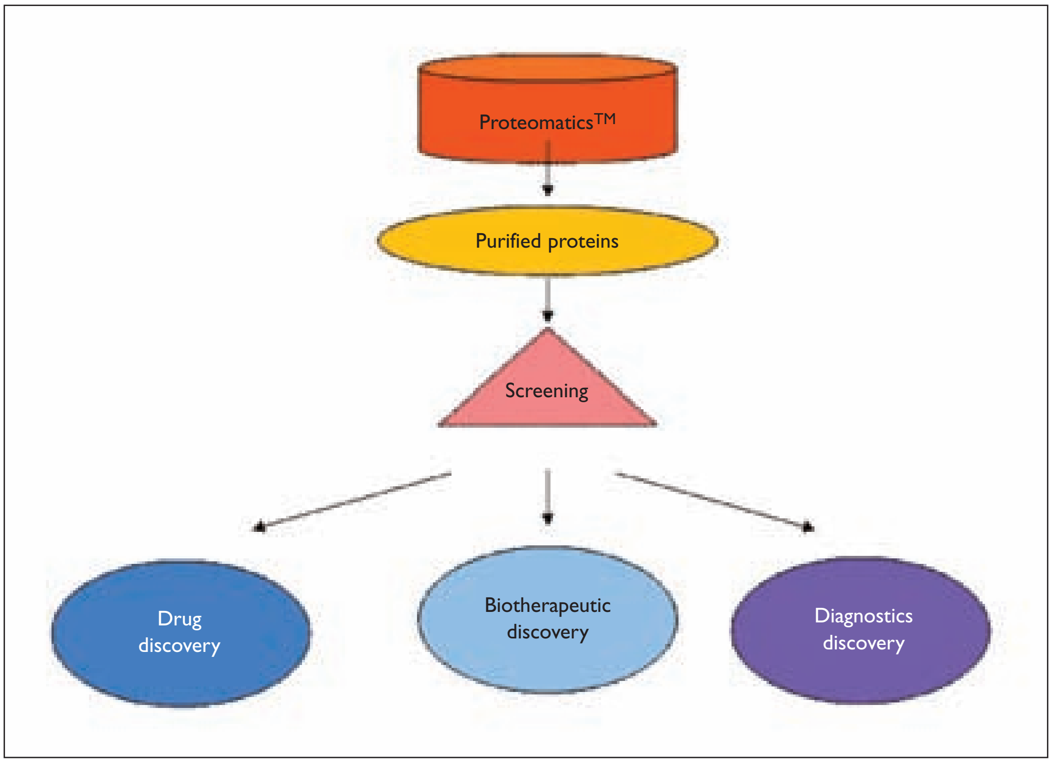


Figure 2
This schema indicates the flow of information from the proteomatics repository to the generation of isolated proteins at a high level of purity to their deployment in screening campaigns. The screening technologies used can be diverse, leading to a wide variety of new applications

Proteomatics

Proteomatics is the process of collating proteomics data and integrating it into a form amenable for exploitation as a knowledge resource in drug discovery, therapeutics and diagnostics. This approach embodies the hierarchical 'data – information – knowledge' paradigm that seems to fit the nature of the central dogma so well. DNA contains the digital database, RNA becomes the information messenger and proteins the functioning embodiment of the message.

To some extent this is a misleading paradigm, since it does not take into account the inherent lack of understanding we have of the processes of genomic expression, mediated through the feedback mechanisms of proteins themselves. Nevertheless, the starting point is the collation of data from a host of proteomics technologies.

This is followed by linking of data and integration of information by means of database technologies, including SWISS-PROT or higher level gene family databases (such as PRINTS or TargetBase®). The final stage is the development of knowledge bases oriented towards individual commercial application areas (eg drug, therapeutic protein or diagnostics discovery).

In order to be successful, proteomatics must collate data from both public domain resources (such as SWISS-PROT, EMBL, GenBank, PubMed, structural databases, image analysis databanks and correlative searching servers) together with proprietary experimental data derived from traditional purification protocols, affinity purification methodologies, separation technologies and analytical processes for complex mixtures. At the information integration level, relevant data items must be linked together even if those items reside in physically separate databases. In this way the power of modern search engines to trawl multiple data warehouses can be exploited and made more efficient.

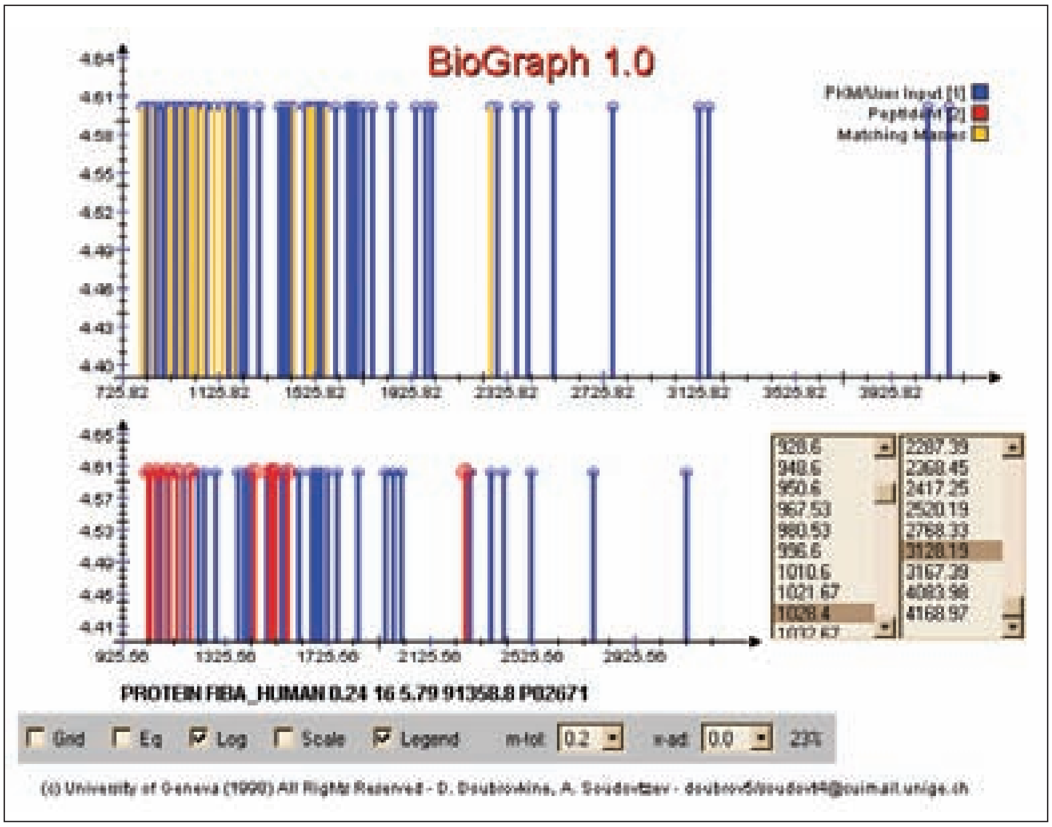
Thus, a new database view is generated that is a synthesis of relevant views from the underlying repositories and tools. For example, Purely Proteins, as a data content provider, is populating the ProteoBase™ component of its Proteomatics™ information architecture, which builds on the target gene family catalogue known as TargetBase®, which already provides a genomic view to support discovery applications.

Software products

Most current products from the software vendors tend to focus in the areas of gel imaging and analysis, correlative database searching, visualisation, database solutions and integrated discovery systems. Visualisation is another important area. A selection of companies and offerings are summarised in **Table 2** in a simple checklist format.

Table 2: A brief list of commercial software vendors and some academic providers of solutions for the proteomics market. For this analysis, the market has been segmented five ways. Some companies have products that cross boundaries into other areas of technology not represented in the table. Many companies are rapidly expanding their coverage and may expand into new areas in the near future

	GEL IMAGING	CORRELATIVE DATABASE SEARCHING	VISUALISATION TOOLS	DATABASE SOLUTIONS	INTEGRATED DISCOVERY PRODUCT
Accelrys				✔	✔
Amersham Biosciences	✔				
Bio-Rad Laboratories	✔				
Decodon	✔				
EMBL		✔		✔	
ExPASy	✔	✔		✔	
GeneBio	✔				
GeneData	✔				
Genomic Solutions	✔				
InforMax					✔
Lion Bioscience				✔	
Matrix Science		✔			
MDS Proteomics		✔			
Nonlinear Dynamics	✔				✔
OmniViz			✔		
Proteome Systems	✔				
Proteometrics		✔			
Rockefeller University		✔			
Scanalytics	✔				
Scimagix	✔				
Scripps Research Institute		✔			
Spotfire			✔		
UK HGMP-RC		✔			



```
928.6      2287.39
940.6      2368.45
950.6      2417.25
967.53     2520.19
980.53     2768.33
996.6      3128.19
1010.6     3167.29
1021.67    4083.98
1028.4     4168.97
1032.67
```

PROTEIN FIBA_HUMAN 0.24 16 5.79 91358.8 P02671

☐ Grid ☐ Eq ☑ Log ☐ Scale ☑ Legend m-tol 0.2 n-ad 0.0 23%

(c) University of Geneva (1990) All Rights Reserved - D. Doubrovkine, A. Soudovtzev - doubrov5/soudov4@cuimail.unige.ch

Figure 3
A BioGraph display of the results of a PeptIdent search performed on the ExPASY website in September 2002. The upper graph is a complete map of the peptide fragment masses from Human Fibrinogen A (SWISS-PROT code FIBA_HUMAN). The supplied masses are shown in blue and the yellow lines indicate matches with the PeptIdent search. The lower graph is a zoomed in version focusing on the masses between 1028.4 and 3128.19 daltons

In order to effectively cover the proteomatics space we need access to gel imaging and analysis technologies, mass spectroscopic analysis and database searching tools, together with databasing and visualisation technologies.

Gel imaging

Nonlinear Dynamics's system, Phoretix for example, has mature capabilities that match the requirements of 2D gel analysis on a routine basis. Spots on the gel matrix can be identified, manipulated and compared with the output in a variety of formats, including interfacing to a spot-picking robot. A version for use with array spot detection is available. XML output is also an option, which makes integration of output into the web environment for the company intranet much more feasible. The TotalLab product is the Nonlinear Dynamics approach to an integrated discovery product, again with a focus very much on image analysis.

Correlative database searching

This is an area that has benefited from substantial academic research effort and the freely accessible sites at ExPASY and the UK HGMP Resource Centre provide links to accessible commercial sites too. An example of fibrinogen masses determined by mass spectroscopy and entered into the PeptIdent fragment search engine on the ExPASY website produces the output shown in **Figure 3**.

Databasing

The storage and effective retrieval of data is one of the least understood areas of bioinformatics. Even after 30 years of research and development, it is hard to have a database accepted by scientists who run research programmes. This is normally because they do not want to lose control of their data. They have total control while it remains in a spreadsheet on their own personal computer. Once it is in the corporate database they perceive, rightly or wrongly, that control and ownership is lost. Coupled with this is the fact that current trends favour relational databases, although designing relational schemas for biological data is often fraught with difficulty. The result tends to be a database that is difficult to query in a flexible manner.

With large comparative datasets, clearly a robust database system must be used to maintain the integrity and security of the data. This applies as well in banking as in proteomics but that does not mean that a database suitable for maintaining customer accounts is the right vehicle for studying the expression of proteins in a multi step signalling

Table 3: A selection of companies providing technology or data content for proteomics. This is a survey based on publicly available information mostly gleaned from websites and news releases. These companies each impact a specific area of the proteomics commercial space. **Key: 1** Through acquisition of INH Technologies; **2** Through collaboration with Neogenesis; **3** Through acquisition of PanVera

COMPANIES			PRODUCTS		
	PROTEOMICS DATA PRODUCED	COMMERCIAL DATABASE AVAILABLE	PROTEIN PURIFICATION SERVICES	DRUG DISCOVERY	PROTEIN THERAPEUTICS DISCOVERY
Affinium Pharmaceuticals	✔			✔	
AxCell Biosciences		✔			
Celera	✔	✔		✔	✔
Cellzome	✔		✔		
Ciphergen	✔			✔	
CPC Biotech AG	✔			✔	
Curagen		✔		✔	✔
Human Genome Sci		✔		✔	✔
Hybrigenics	✔				✔
Incyte Genomics		✔		✔	
MDS Proteomics	✔			1	
Myriad Genetics	✔	✔			
Neogenesis	✔			✔	
Oxford Glycosciences	✔	✔		2	
Proteome Sciences	✔				
Purely Proteins	✔	✔	✔	✔	✔
Vertex	✔		3	✔	
Wita Proteomics	✔				

cascade in diseased and healthy liver cells. Careful thought must be given to database solutions bearing in mind the application to which the data will be put, ie the three application areas mentioned earlier. Databases oriented towards answering questions that impinge further up the value chain will have greatest impact in the long run.

Visualisation

Naturally, visualisation occurs at each stage in the proteomatics process. Gel imaging and mass spectra are both examples of direct visualisation of data. However, when it comes to analysing the results of a comparative proteomics experiment, powerful tools for visualisation of clusters can be useful for exploration. OmniViz is one such tool that is currently gaining in popularity. Many scientists are already using Spotfire to visualise tabular datasets as an adjunct to Excel spreadsheet graphics. However, for any visualisation tool that comes packaged with standard algorithms, some customisation is required to make sure that the relevant analytical techniques are available within the visualisation environment. Training is a key issue in using these techniques to gain real insight into the data.

Content providers

Most biotechnologies are cumulative: information gained using one technology can be used as a springboard for applications in another. This is particularly true for proteomics, where genomics has provided a rich foundation on which to build an understanding of this newly emerging area. Mature genomics companies have not been slow to realise this and many of the smaller proteomics companies have been the subject of early acquisition or merger, a trend which is extending further – into the structural biology and drug discovery technology zones. The corporate landscape in this area will no doubt evolve rapidly.

Thus, there are two breeds of proteomics company: those (such as Celera and Incyte) which have evolved from genomics into proteomics companies as the field has matured, and those (such as MDS Proteomics and OGS) which have grown up around the various technologies for characterisation and analysis of proteins. A selection of these companies, together with their areas of expertise, is shown in **Table 3**.

The commercial challenge is now to provide a framework within which to integrate the widely disparate information sets available within these specialised companies.

Conclusions

Proteomatics is a newly emerging information approach which aims to integrate the fields of proteomics and bioinformatics. To implement a resource that covers the whole potential space of the technology, new modular database solutions need to be developed, and more finely focused datasets generated, so that the direct impact of these technologies on drug discovery, protein therapeutic discovery and diagnostic discovery can be realised and the commercial opportunities reduced to practice. DDW

__Dr David Parry-Smith__ is a founder of Purely Proteins and Chief Information Officer. He brings substantial experience of large pharma bioinformatics and genomics database projects. David was a member of the founding team of Cambridge Drug Discovery in 1997. He writes and lectures widely in bioinformatics and knowledge management.

__Dr David Bailey__ is a founder of Purely Proteins and Chief Executive Officer. David has significant experience in the pharmaceutical industry, holding senior management positions with three large companies before founding De Novo Pharmaceuticals in 1999. His mission is to use genomics information to increase the speed and accuracy of the drug discovery process.

Reference

[1] Pennington, SR and Dunn, MJ. Proteomics: From protein sequence to function. BIOS Scientific Publishers Limited, Oxford, UK (2001).

ADVERTISEMENT INDEX

Affymetrix, Inc	32	Gene Logic Inc	30-31	MDS Proteomics	78
Applied Biosystems	63	Genomic Solutions, Inc	8	Molecular Devices Corporation	16
Asterand, Inc	37	Gilson Inc	IFC-3	PanVera Corporation	OBC
Beckman Coulter, Inc	15	Golden Helix, Inc	53	Perkin Elmer Life Sciences	6,45
Biacore International SA	26	IBC USA	29	Purely Proteins Ltd	43
Cellomics Inc	69	IBC Europe	23,IBC	Scientific Software, Inc	70
Ciphergen Biosystems, Inc	24	Informax, Inc	4	Thermo Finnigan Corporation	21
DxS Ltd	46	Luminex Corporation	58		